An Evaluation of Genetic Distances for Use With Microsatellite Loci

David B. Goldstein,* Andres Ruiz Linares,†,‡ Luigi Luca Cavalli-Sforza‡ and Marcus W. Feldman*

**Department of Biological Sciences, Stanford University, Stanford, California 94305, †Universidad de Antioquia, Facultad de Medicina, Centro de Investigaciones Medicas, Medellin, Colombia and ‡Department of Genetics, Stanford University, Stanford, California 94305*

Manuscript received June 22, 1994
Accepted for publication October 5, 1994

ABSTRACT

Mutations of alleles at microsatellite loci tend to result in alleles with repeat scores similar to those of the alleles from which they were derived. Therefore the difference in repeat score between alleles carries information about the amount of time that has passed since they shared a common ancestral allele. This information is ignored by genetic distances based on the infinite alleles model. Here we develop a genetic distance based on the stepwise mutation model that includes allelic repeat score. We adapt earlier treatments of the stepwise mutation model to show analytically that the expectation of this distance is a linear function of time. We then use computer simulations to evaluate the overall reliability of this distance and to compare it with allele sharing and Nei's distance. We find that no distance is uniformly superior for all purposes, but that for phylogenetic reconstruction of taxa that are sufficiently diverged, our new distance is preferable.

STUDIES of phylogenetic relationships among very closely related species are often hampered by a lack of variation. For example, in their study of mitochondrial DNA variation in the Lake Victoria flock of East African cichlids, MEYER *et al.* (1990) found no variation in a 363-bp region of the cytochrome *b* gene and an average of two or three substitutions separating species in 440 bp of the control region. The estimation of relationships among such closely related taxa, or the estimation of relationships within a species, would be easier if faster evolving characters were used. Because of their exceptionally high mutation rate, microsatellites may prove more informative for working out relationships among such closely related species, as well as among subpopulations of a single species (BOWCOCK *et al.* 1994).

Microsatellites are a special class of tandem repeat loci that involve a base motif of 1–6 bp repeated up to ~100 times (TAUTZ 1993). The few tandem repeat loci that have been studied show exceptionally high mutation rates, with minimal rates as high as 10^{-3} (JEFFREYS *et al.* 1988; KELLY *et al.* 1991) or 10^{-4} (LEVINSON and GUTMAN 1987; HENDERSON and PETES 1992). Because of this and the generally large number of alleles available, these loci have been extremely useful in DNA fingerprinting (JEFFREYS and PENA 1993; QUELLER *et al.* 1993), linkage analysis (TODD *et al.* 1991; DIETRICH *et al.* 1992) and more recently in the reconstruction of human phylogeny (BOWCOCK *et al.* 1994).

Although microsatellites may prove to be more useful than classical polymorphisms (or sequence data) for assessing population structure and determining the relationships among very closely related species, they may be less informative for more distantly related taxa. This is because the range of variation in number of repeats, while large, is ultimately restricted (BOWCOCK *et al.* 1994). Therefore, after sufficient time has passed, any distance applied to these loci will reach a maximal value.

Although a large number of evolutionary distances could conceivably be applied to microsatellites (see below), there has been relatively little theoretical evaluation of inferences and their reliability (but see CHAKRABORTY and JIN 1993). Furthermore, alleles at some of these loci are thought to evolve by a stepwise mutation process, in which an allele mutates up or down by a small number of repeats (SCHLÖTTERER and TAUTZ 1992). Therefore, estimators of population parameters based on the infinite-alleles model, for example, are unlikely to apply to microsatellite loci. Although this stepwise mutation model is consistent with the distribution of alleles at microsatellite loci (SHRIVER *et al.* 1993; VALDES *et al.* 1993), it is probably not applicable to minisatellite loci (SHRIVER *et al.* 1993). Furthermore, DI RIENZO *et al.* (1994) provide evidence that a strict (single-step) stepwise mutation model may not be sufficient to account for allele frequency distributions at microsatellite loci.

In this paper we first derive a distance measure that is linear with time when applied to loci undergoing a strict stepwise mutation process with no constraint on allele size and then use computer simulations to evaluate the reliability of this and other distances. We con-

Corresponding author: David B. Goldstein, Department of Biology, The Pennsylvania State University, 208 Mueller Laboratory, University Park, PA 16802-5301. E-mail: david@kimura.stanford.edu

sider both the reliability of phylogenetic reconstruction and the reliability of inferences about the populations to which individuals belong.

A LINEAR GENETIC DISTANCE FOR MICROSATELLITES

Here we adapt earlier treatments of the stepwise mutation model to obtain an evolutionary distance whose expectation is linear with time. The reliability of an evolutionary distance depends both on its expectation and variance, but a linear expectation is desirable if it does not entail too large an increase in the variance. OHTA and KIMURA's (1973) stepwise mutation model was originally applied to changes in the charge state of proteins as inferred from electrophoretic mobility. More detailed mathematical and statistical analyses of this model, including the possibility of two-step mutations, were made by MORAN (1975), WEHRHAHN (1975), BROWN *et al.* (1975), WEIR *et al.* (1976) and others.

We consider here only the strict stepwise mutation model, in which an allele with i repeats mutates to each of $i-1$ and $i+1$ repeats with probability $\mu/2$. Assume a population of N diploid individuals. MORAN (1975) showed that with multinomial sampling, the probability distribution of $n_i(t)$, the number of gametes carrying an allele with i repeats at time t, does not converge as t grows large. Similarly, the mean number of repeats, $(2N)^{-1}\Sigma_i\, in_i(t)$, does not converge (although the expected value of this mean does not change with time). The variance in repeat number, given the initial conditions, however, does converge. Thus, although the average of the number of repeats never reaches an equilibrium value, the "cloud" around the mean retains constant variance as the mean position wanders. MORAN (1975) also showed that the random variables $C_k(t) = (2N)^{-2}\Sigma_i\, n_i(t)\, n_{i+k}(t)$ converge. For example, $E[C_0]$ approaches $(1+2\theta)^{-1/2}$, where $\theta = 4N\mu$, a result also derived by OHTA and KIMURA (1973). Its reciprocal is the effective number of alleles.

If D_0 is defined as the average squared difference in repeat numbers for two alleles drawn from the same population, then direct application of MORAN's (1975) results yields the limiting expectation, $E(D_0) = 2(2N-1)\mu$. Similarly, define D_1 as the average squared difference in repeat numbers for two alleles drawn one each from different populations isolated τ generations in the past. For convenience we will refer to a distance based on D_1 as the average squared distance. In the APPENDIX we show that the expectation of D_1 is a linear function of τ. Specifically,

$$E[D_1(\tau)] = 2(2N-1)\mu + \tau 2\mu. \qquad (1)$$

SLATKIN (1995) has also derived very similar expectations using coalescent theory. He found that the average squared difference between alleles is $2\mu\sigma_m^2\bar{t}$, where $\bar{t}$ is the expected coalescence time between the alleles, μ is the mutation rate and σ_m^2 is the variance of the distribution of mutations. Within a single population $\bar{t}$ is $2N$ (HUDSON 1990). This gives $E(D_0) = 4N\mu$ for the strict stepwise mutation model. The expected coalescence time for alleles drawn from each of two populations separated τ generations ago is $2N+\tau$ (SLATKIN 1995), from which $E(D_1)$ may be derived. These results agree with those derived above upon substitution of $2N$ for $2N-1$. Note that Slatkin's analysis allows mutations of more than one repeat unit in the stepwise mutation model.

INFERENCES BASED ON VARIATION AT MICROSATELLITE LOCI

Estimation of D_1: If it is known *a priori* to which population each individual belongs, the phylogenetic relationships among the populations can be inferred using an estimate of D_1 based on the sampled individuals. An obvious estimator of D_1 can be written in terms of the repeat scores of two sampled alleles i, i'. Namely,

$$\Delta_{ii'} = (i-i')^2. \qquad (2)$$

It is easy to see that the average ($\bar{\Delta}$) of Δ between all alleles sampled, one from each population, is an unbiased estimator of $E[D_1]$. Denote E_s as the expectation under random sampling from the two populations, then the expectation of ($\bar{\Delta}$) is $E_s[\bar{\Delta}] = E_s\Sigma_i\Sigma_{i'}(i-i')^2 f_i f_{i'}$, where the sum is over all i, i', and f_i, $f_{i'}$ are, respectively, the frequencies of alleles i and i' in the samples from the first and the second population. Thus, because sampling is independent in the two populations, we have $E_s[\bar{\Delta}] = \Sigma_i\Sigma_{i'}(i-i')^2 E_s[f_i]E_s[f_{i'}] = \Sigma_i\Sigma_{i'}(i-i')^2 F_i F_{i'} = D_1$ where F_i, $F_{i'}$ are, respectively, the parametric frequencies of alleles i and i' in the first and second population. Therefore, $\bar{\Delta}$ is an unbiased estimator of D_1. Thus, the expectation of $\bar{\Delta}$ should be a linear function of the time since the populations were isolated.

Phylogeny reconstruction: We confirmed the linearity of $\bar{\Delta}$ using computer simulation and also compared its behavior with two other distances: allele sharing (D_{AS}), which has recently been used to infer human phylogenetic relationships (BOWCOCK *et al.* 1994), and Nei's distance (NEI 1972), denoted D_S. In Figure 1 the mean behaviors of $\bar{\Delta}$, D_{AS} and D_S are compared. The figure reports the results of 100 independent simulations, all using the same parameters (N, μ). For convenience we will refer to the generation time in comparing the behavior of the distances, but it should be noted that such references are specific to the particular values of N and μ used in the simulation. Notice that after sufficient time has passed (here about 1000 generations), both D_{AS} and D_S are beginning to asymptote,

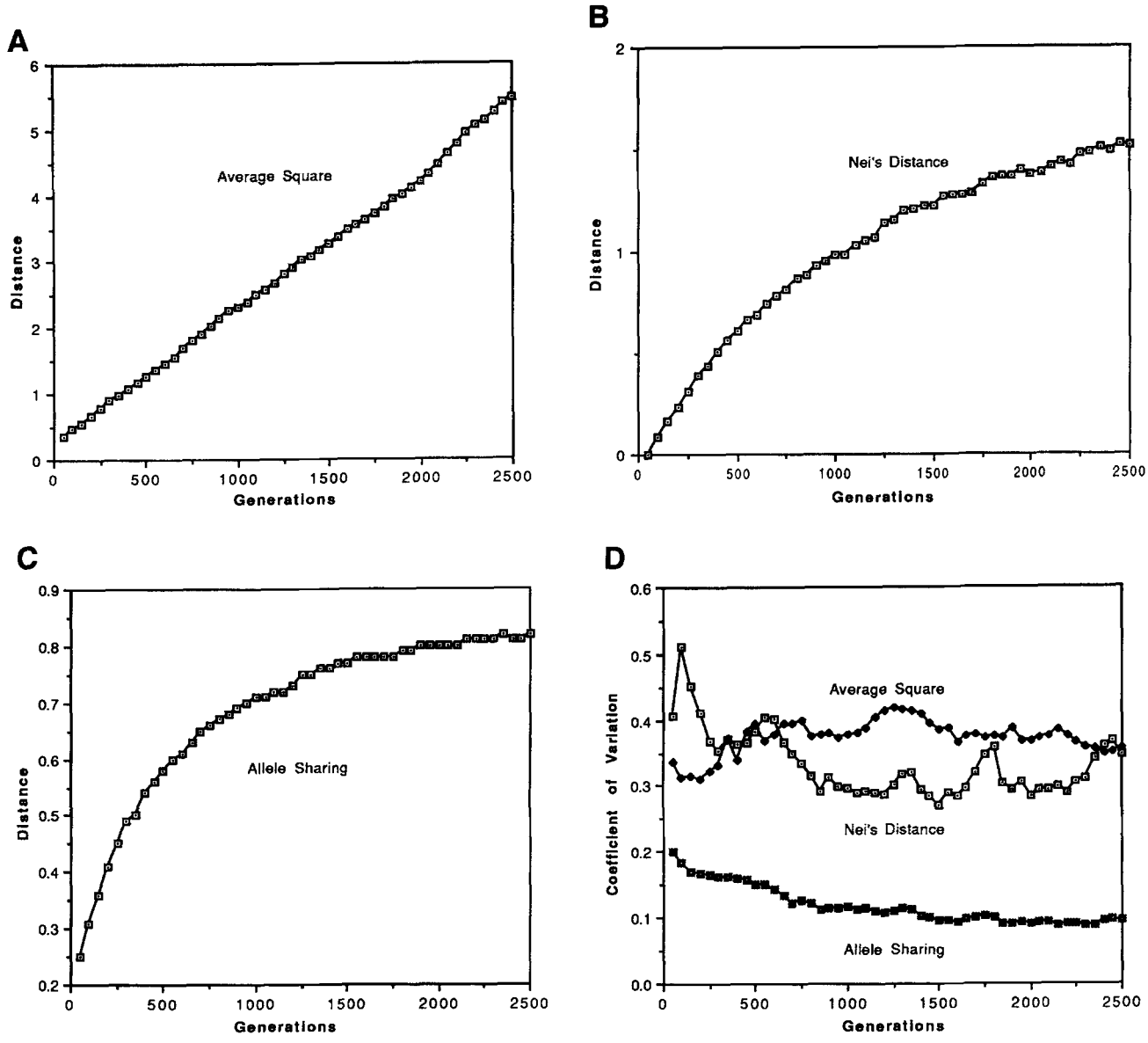

FIGURE 1.—Means and variances of the distances as a function of time. To determine the mean of the distance estimates as a function of time, we started with two identical populations with an equilibrium variance of repeat number. This equilibrium was reached by iteration. We simulated the independent evolution of these populations for 2500 generations. The haploid population size of each taxon was 200 and an allele in state i mutated to state $i + 1$ and $i - 1$ with probability 0.0005 each per generation. The next generation was formed by random sampling of the previous generation. We calculated each distance ($\bar{\Delta}$, D_{AS}, D_S) every 100 generations. D_{AS} was calculated based directly on the haploid populations as 1 − (the average number of shared alleles). That is, $D_{AS} = 1 - (1/2N)^{-2} \Sigma_i \Sigma_i' I(i, i')$, where the first and second sums are over all alleles in the first and second population, and $I(i, i')$ is an indicator variable that equals 1 if the alleles are the same and zero if they are not. The formula for $\bar{\Delta}$ is given in the text, and the formula for D_S (Nei's "standard" distance) can be found in NEI (1972). Figure 1, A–C show, respectively, the average values of $\bar{\Delta}$, D_S and D_{AS} among 100 independent simulations. Figure 1D shows the coefficients of variation for the three measures.

but $\bar{\Delta}$ remains linear. A log transformation of D_{AS} improves its linearity only slightly, making it similar to D_S. Thus, we will not consider this further here.

Because the variance of an estimator also influences its performance, we estimated the coefficients of variation of the distances (Figure 1D). Notice that the coefficient of variation of D_{AS} is always smaller than that of $\bar{\Delta}$ and D_S. Thus, D_{AS} is the superior distance with respect to variance but inferior with respect to expectation. One criterion that can be used to determine which distance is superior is to compare the slope of the mean to the standard deviation at each point in time (TAJIMA and TAKEZAKI 1994). This shows that $\bar{\Delta}$ is superior after ~100 generations (data not shown). Trees that include more distantly related taxa, therefore, should tend to be more accurately reconstructed by $\bar{\Delta}$.

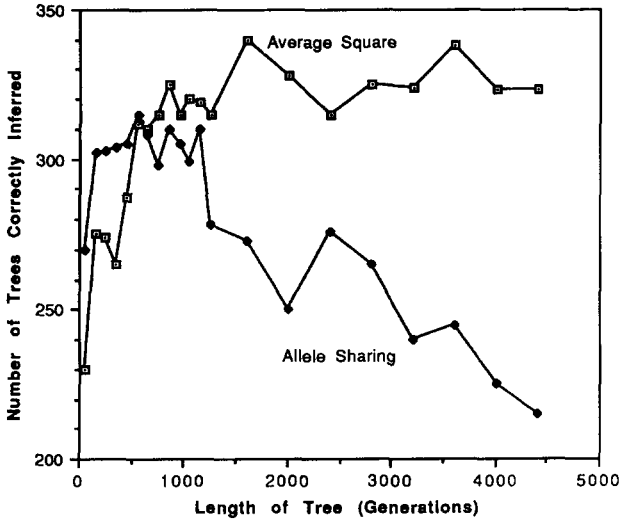

FIGURE 2.—Reliability of the distances in recovering the correct phylogeny. To test how the two distance measures influence the reliability of phylogenetic reconstruction, we simulated the stepwise mutation process along a three-taxon tree of variable length and inferred the phylogenetic relationship among the taxa at the end of the simulation using the UPGMA algorithm with each distance estimate. The total length of the tree varied from 50 to 4400 generations, the haploid population size of each taxon was 200, and the number of loci was 15. The single speciation event occurred at the exact midpoint of the tree. We ran 400 independent simulations along each of the trees. The curves represent the number of times out of 400 that the correct topology was inferred.

It is important to determine the expected difference in performance between D_{AS} and $\bar{\Delta}$ for different types of trees. We compared the overall reliability of D_{AS} and $\bar{\Delta}$ by simulating evolution along a three-taxon tree of variable length. Figure 2 shows that for trees of a total length shorter than ~300 generations, D_{AS} reconstructs phylogenies more accurately than $\bar{\Delta}$. From ~500 generations on, however, $\bar{\Delta}$ is progressively more reliable than D_{AS}.

Note from Equation 1 that, except for a constant factor independent of time, $\bar{\Delta}$ is not a function of the population size. Therefore, unlike distances such as F_{ST} that measure the differentiation caused by sampling and not mutation, the appropriate measure of time for $\bar{\Delta}$ is $\mu\tau$, not $N\tau$.

Revealing cryptic population structure: For some data sets it may not be possible to determine membership of individuals in populations before genetic analysis is carried out. For example, we might be interested in testing whether a population that is not obviously partitioned into isolated subgroups is in fact genetically structured because of behavior (*e.g.*, BOWEN *et al.* 1993) or because of recent admixture among a set of previously distinct populations (*e.g.*, BOWCOCK *et al.* 1994).

When the proper assignment of individuals to populations is not known, the mean squared differences among alleles in pairs of individuals can be split into two components, giving information about D_1 and D_0. Let the alleles in one individual be i_1, i_2 and those in the other be j_1, j_2. The total squared difference among the four alleles in these two individuals is $V_T = 2(i_1 - i_2)^2 + 2(j_1 - j_2)^2 + (j_1 + j_2 - i_1 - i_2)^2$. The within- and between-individual components of this sum of squares are $(i_1 - i_2)^2 + (j_1 - j_2)^2$ and $(i_1 - j_1)^2 + (i_1 - j_2)^2 + (i_2 - j_1)^2 + (i_2 - j_2)^2$, respectively. The within-individual component is an estimate of D_0 (although not unbiased), because alleles within an individual must come from the same population (ignoring admixture). At a single locus there is a total of N (total population size) observations that can be used to estimate D_0. These may not all estimate the same value of D_0, however, because the different subpopulations may have different population sizes.

The between-individual component of the sum of squares will reflect either D_0 or D_1, depending on whether the two individuals come from the same or different populations. A matrix of these between-individual squared differences, therefore, will have elements each of which is an estimate of either D_0 or D_1. If these estimates are sufficiently different, a clustering program will group individuals into their correct populations. This is the basis for the approach taken by BOWCOCK *et al.* (1994), who used a distance measure based on the proportion of shared alleles between individuals (see also STEPHENS *et al.* 1992; CHAKRABORTY and JIN 1993). They found that trees of individuals based on this distance are structured into taxonomic units that correspond well with the geographic origins of the individuals. This suggests that microsatellite loci can be used to assign individuals to the populations from which they come.

We simulated evolution on a three-taxon tree to determine whether $\bar{\Delta}$ or D_{AS} more accurately assigns individuals to their correct populations. Figure 3 shows that over all trees evaluated D_{AS} makes this assignment more accurately. Because $\bar{\Delta}$ is more accurate over long periods of time, this is at first puzzling. A possible explanation may involve the method of assignment of individuals to populations. Denote by D_W, the expected value of either distance for two individuals drawn from the same population and by D_B, the same for individuals drawn from different populations. (If there are more than two subpopulations there will be a series of D_B's, one for each pair of populations.) For the individuals to be placed into taxonomic groups corresponding to the populations from which they come, it is necessary and sufficient that D_B be distinguishable from D_W. It is not necessary that any of the different values of D_B (representing individuals drawn from different pairs of populations) be distinguishable. Therefore, the fact that D_{AS} asymptotes more quickly than D_1 does not interfere with the assignment of individuals to populations, and its greater precision (lower variance) makes it the

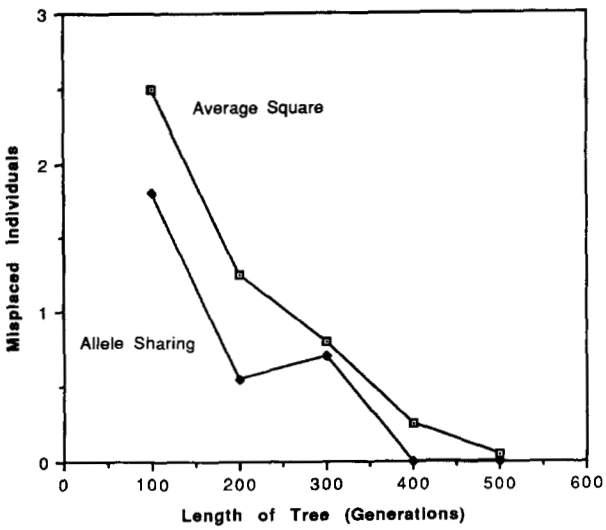

FIGURE 3.—Reliability of assignment of individuals to populations. To test how well $\bar{\Delta}$ and allele sharing assign individuals to populations, we simulated evolution along a three-taxon tree as described above but with 75 haploid individuals of each taxon. The trees ranged in length from 100 to 600 generations, and 100 replications were run for each tree length. At the end of the simulation, we assumed that the assignment of individuals to populations was unknown. We then used either $\bar{\Delta}$ or allele sharing (see text) to build a tree of individuals. The structure of this tree was then studied, and the average number of individuals incorrectly assigned (out of 225) among the 100 replications was calculated.

superior distance for this purpose. Of course, the arrangement of the populations that results from D_{AS} may be incorrect, but individuals coming from the same population will tend to cluster near one another regardless of the arrangement of the populations within the tree.

PRACTICAL CONSIDERATIONS

Constraints on the maximum number of repeats: The treatment above assumed that alleles can mutate to arbitrarily large or small repeat scores. In reality the number of possible repeat scores is restricted, and several lines of evidence suggest that this limit is fairly strict. BOWCOCK *et al.* (1994) found that the variance in repeat score is approximately the same within humans and within primates. If there were no restriction, one would expect the greater evolutionary distance among primates to lead to much greater differences in repeat scores. Additionally, of the 20 loci typed in humans and the 10 also typed in the other primates (unpublished data), all but one has a maximal repeat score under 100. The demonstrated connection between large repeat scores and hypermutability (KUNKEL 1993; STRAND *et al.* 1993) suggests a mechanism for the constraint on repeat score.

The length of time during which D_1 is (approximately) linear as a function of the maximum number of alleles possible, R (which is also the range of the repeat score), can be approximated as follows. Assume that each of two isolated populations has only one type of allele. Then, it is possible to calculate the average value of D_1 after the isolated populations have reached maximal divergence. In this case the repeat score of the single allele in each population is randomly drawn from the possible R values. Therefore, the joint probability distribution for the process of sampling an allele in state i from one population and an allele in state j from the other population is given by $p(i, j) = 1/R^2$. The maximum divergence possible as a function of R, denoted $\vartheta(R)$, is the expectation of $(i - j)^2$ and is given by

$$\vartheta(R) = \frac{\Sigma_i \Sigma_j (i - j)^2}{R^2} = \frac{R^2 - 1}{6}. \qquad (3)$$

Substituting $\vartheta(R)$ for $D_1(t)$ in Equation 1 and solving for τ yields the amount of time it takes D_1 to reach the value it takes at maximal divergence. This time should approximate the duration of linearity of D_1 and is given by

$$\tau_R = \frac{\vartheta(R)}{2\mu} - 2N + 1. \qquad (4)$$

Figure 4 shows the results of computer simulations similar to those shown in Figure 1, except that the total number of alleles possible is restricted to 6 (Figure 4A) and 10 (Figure 4B). First, note that $\bar{\Delta}$ asymptotes at ~6 and 17 for the respective simulations. This compares nicely with predictions of $\vartheta(6) = 5.8$ and $\vartheta(10) = 16.3$. Thus, the approximation for the maximum value of D_1 given in Equation 3 appears to be quite accurate under these conditions.

In Figure 4 a hyperbolic function has been fitted to the simulation results. We take the point at which linearity is lost to be that point on the hyperbola where a tangent has slope halfway between the initial slope (based on Equation 1) and the final value (zero). This method produces estimates of 1710 and 5900 for the approximate times that the simulated values of $\bar{\Delta}$ stopped their linear increase with time, somewhat lower than predicted by Equation 4 (2800 and 8100, respectively). This is probably due to the fact that the simulation reports the average values of a nonlinear function, from which we wish to infer the value of its argument (time). Because the average value of a nonlinear function is not the same as the function applied to the average of its argument, the method used in Figure 4 provides a biased estimate of the time at which the slope is halfway between its initial and final values. Nonetheless, it appears that Equation 4 provides a useful approximation of the range of linearity of D_1. For comparison, Figure 4C shows the dynamic behavior of D_{AS} for $R =$

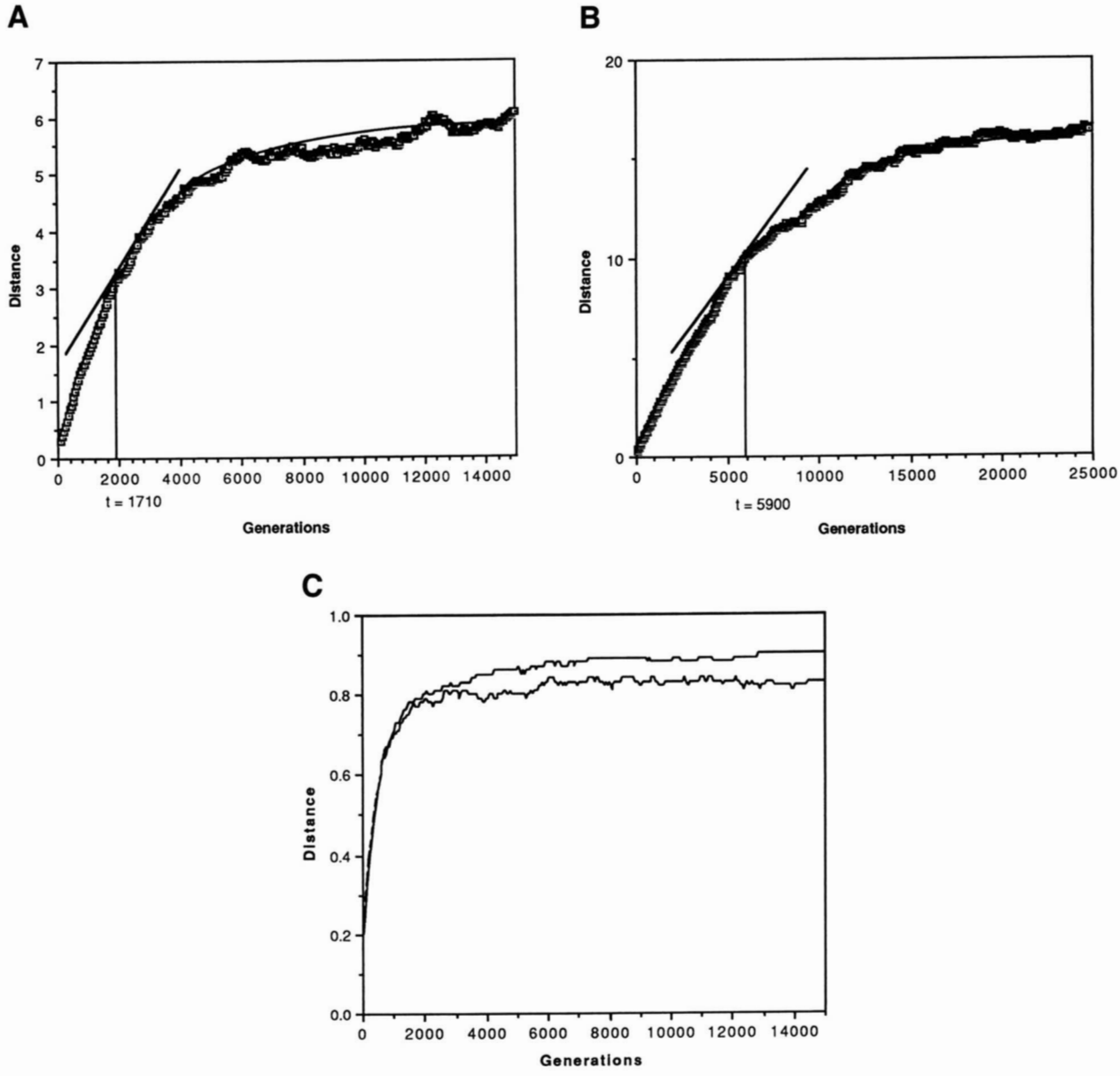

FIGURE 4.—Duration of linearity of D_1. To determine the duration of linearity of D_1, we ran simulations identical to those described in Figure 1, except that the allelic range was restricted to 6 (A) or 10 (B). The points show the results of simulated evolution in a single population with 160 loci. (The results are nearly identical if fewer loci are used in replicate populations). The curve is the best-fitting hyperbola of the form $a - b/(c + dt)$, where t is the number of generations, and a, b, c, d are the parameters to be fit. The tangent to the hyperbola is a line with slope equal to (initial slope − final slope)/2. The value of time at this point on the hyperbola is the estimate of the length of time during which linearity is maintained (see text). C shows the behavior of D_{AS} for each of the conditions. The higher curve (broken) is for 10 alleles.

6 and $R = 10$. Unfortunately, because we do not have a dynamic for D_{AS}, we cannot use the approach described above to estimate the time at which D_{AS} reaches its asymptotic value. It is important to note, however, that for $R = 10$, D_{AS} clearly asymptotes earlier than does D_1. Thus, as long as R is not very small, we can conclude that D_{AS} asymptotes well before D_1.

Because the duration of linearity is expected to grow as the square of R, loci with slightly larger values of R may be informative over much longer times than those with smaller values. Table 1 shows approximate observed ranges of the 30 loci typed in humans (unpublished data) and the predicted duration of linearity based on those ranges. The values must be regarded only as approximate because the primers generally do not exactly enclose bases making up the tandem repeat but include a variable number of nonrepeated bases on either side of the microsatellite.

Two ranges were calculated for each locus. The first is simply the difference between the maximal and mini-

TABLE 1

Estimated range of linearity of 30 microstatellite loci typed in humans

Locus	Range[a]	Expected duration of linearity of D_1[b]
084XC5	10 (48)	0 (171,917)
D13S126	9 (56)	0 (241,251)
D13S119	14 (69)	0 (376,667)
D13S118	7 (99)	0 (796,667)
D13S125	15 (79)	0 (500,001)
D13S144	9 (98)	0 (780,251)
1523	5 (91)	0 (670,001)
ACTC	14 (48)	0 (171,917)
D15S171	8 (61)	0 (290,001)
D15S169	11 (80)	0 (513,251)
D13S133	35 (93)	82,001 (700,667)
D13S137	12 (60)	0 (279,917)
D13S227	16 (81)	1,251 (526,667)
FES	12 (82)	0 (540,251)
GABRB3	10 (99)	0 (796,667)
D13S192	14 (58)	0 (260,251)
D13S193	11 (74)	0 (436,251)
HLIP	7 (86)	0 (596,251)
D15S98	17 (86)	4,001 (596,251)
D15S97	15 (91)	0 (670,001)
D15S100	15 (65)	0 (332,001)
D15S101	12 (68)	0 (365,251)
D13S115	9 (89)	0 (640,001)
D15S95	9 (74)	0 (436,251)
D15S108	12 (80)	0 (513,251)
D15S11	9 (41)	0 (120,001)
D15S102	11 (58)	0 (260,251)
D15S117	12 (74)	0 (436,251)
D15S148	10 (74)	0 (436,251)
p4-3R	13 (131)	0 (1,410,001)

[a] The first range was calculated as the maximal allele score − minimal allele score. The range in parentheses is the maximal allele score − 2.

[b] Duration expressed as no. of generations.

mal repeat scores at that locus. The second range (in parentheses) assumes that all loci have a minimum repeat score of 2 and calculates the range as the difference between the maximal repeat score and 2. (Two was chosen as the minimum required to trigger the increased rate of mutation associated with tandem repeats.) The ranges were used to predict the duration of linearity, assuming a population size of 10,000 diploid individuals and a mutation rate of 10^{-3}.

Note that there are two sources of error in estimating the allelic range. The first is due to the fact that, unless R is very small, a population will not include all possible alleles. Second, we have available only a sample from the actual population. Both of these errors result in underestimation, which means that the biological upper limits on repeat scores are likely to be substantially larger than those reported here. In addition, the assumed mutation rate is too high for some microsatellites.

Although a number of strong assumptions were made in producing the estimates of the duration of linearity in Table 1, the results suggest that (at least some) microsatellite loci may be useful for resolving distances as far back as several million years. Furthermore, the great variation in allelic range among loci suggests that it might be necessary to select only those loci with many alleles to study more distantly related taxa. This raises the more general question of how best to combine information across loci.

Multiple loci: We developed the average squared distance for a single locus, but its expectation remains linear when it is applied to many loci, even if the mutation rate varies across loci. In this case, the slope of $E[D_1]$ becomes the arithmetic average of the mutation rates across the loci. However, the loci will not be equally informative. More polymorphic loci will provide a more reliable signal over a longer period of time. Ideally, to combine information across loci, a variance weighting approach similar to that described in GOLDSTEIN and POLLOCK (1994) would be desirable. However, deriving the variance of D_1 (and an estimator for this variance) is not trivial, especially if one also considers constraints on repeat score, which are obviously critical for determining a good weighting. In a subsequent paper we will address the problem of weighting. For now, the fact that D_1 remains linear when averaged over loci with different mutation rates seems sufficient grounds to motivate its use on multi-locus data sets.

Details of the mutation process: As pointed out by many authors (HENDERSON and PETES 1992; SCHLÖTTERER and TAUTZ 1992) and emphasized by SLATKIN (1995), the mutation process at microsatellite loci is not memoryless. That is, when a mutation occurs the new mutant is related to the allele from which it was derived. In this case the difference in length between alleles contains phylogenetic information. The statistics developed here and those developed by SLATKIN (1995) were designed to include this information and are equivalent except that Slatkin takes a ratio of combinations of the Ds to eliminate the parameters of the mutation process. This is necessary to estimate the actual value of demographic parameters. The confounding of time and mutation rate does not present a problem for the estimation of phylogenetic relationship, however, because all we need is a linear function of that time. Thus, we use D_1 directly as a phylogenetic distance measure.

In deriving and evaluating the statistics reported here, we assumed a strict stepwise mutation process, which probably does not hold for many microsatellite loci (DI RIENZO *et al.* 1994; SHRIVER *et al.* 1993). This assumption is not a problem with respect to the expectation of our distance. SLATKIN's (1995) results show that the statistic D_1 will remain linear if the mutation model

is relaxed to include mutations of larger effect. Instead of a distance measure with slope $\bar{\mu}$, where $\bar{\mu}$ is the average mutation rate across loci, the distance measure would have slope $\mu\sigma_m^2$. However, the variances of D_1, D_{AS} and D_S will depend on the exact details of the mutation process, and the relative performance of the distances will therefore depend on those details.

In fact, it is clear that the average squared distance will do progressively worse as the mutation model becomes more like the infinite alleles model. Under the infinite-alleles model, we know that Nei's distance is linear and that the average square includes information that is not related to time since common ancestry (allelic repeat score). Thus, the average square must be noisier and will perform worse. For a mixed model, we can assume that as the proportion of single-step mutations is reduced (and the proportion of arbitrary size goes up), the performance of the average square relative to Nei's distance will decline.

A potentially more serious complication is that the mutation process may depend on the repeat score. If this dependence were strong, the results reported here would not be relevant. In particular, it might be inappropriate to consider a repeat score of 2 as a reflecting boundary (WALSH 1987), and, more generally, if the mutation rate depends on the repeat score, D_1 may not be linear.

We thank J. EISEN, J. KUMM, D. POLLOCK, M. SLATKIN and an anonymous reviewer for helpful discussions and/or comments on earlier versions of this manuscript. This research was supported in part by National Institutes of Health grants GM-28016 to M.W.F., GM-20467 to L.C.S. and GM-28428 to L.C.S. and M.W.F. A.R.L. was supported in part by Colciencias and the Programa de Reproduccion, Universidad de Antioquia. A program that calculates various distances for microsatellites is available from Dr. Eric Minch, Department of Genetics, Stanford University. E-mail: Eric@Lotka.Stanford.edu

LITERATURE CITED

BOWCOCK, A. M., A. RUIZ-LINARES, J. TOMFOHRDE, E. MINCH, J. R. KIDD *et al.*, 1994 High resolution of human evolutionary trees with polymorphic microsatellites. Nature **368:** 455–457.

BOWEN, B. W., J. I. RICHARDSON, A. B. MEYLAN, D. MARGARITOULIS, R. HOPKINS MURPHY and J. C. AVISE, 1993 Population structure of loggerhead turtles (*Caretta caretta*) in the northwestern Atlantic Ocean and Mediterranean Sea. Conserv. Biol. **7:** 834–844.

BROWN, A. H. D., D. R. MARSHALL and L. ALBERCH, 1975 Profiles of electrophoretic alleles in natural populations. Genet. Res. **25:** 137–143.

CHAKRABORTY, R., and L. JIN, 1993 A unified approach to study hypervariable polymorphisms: statistical considerations of determining relatedness and population distances, pp. 153–175 in *DNA Fingerprinting: State of the Science,* edited by S. D. J. PENA, R. CHAKRABORTY, J. T. EPPLEN and A. J. JEFFREYS. Birkhäuser Verlag: Basel.

DIETRICH, W., H. KATZ, S. E. LINCOLN, H.-S. SHIN, J. FRIEDMAN *et al.*, 1992 A genetic map of the mouse suitable for intraspecific crosses. Genetics **131:** 423–447.

DI RIENZO, A., A. C. PETERSON, J. C. GARZA, A. M. VALDES, M. SLATKIN *et al.*, 1994 Mutational processes of simple sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA **91:** 3166–3170.

GOLDSTEIN, D. B., and D. D. POLLOCK, 1994 Least squares estimation of molecular distance—noise abatement in phylogenetic reconstruction. Theor. Popul. Biol. **45:** 219–226.

HENDERSON, S. T., and T. D. PETES, 1992 Instability of simple sequence DNA in *Saccharomyces cerevisiae.* Mol. Cell. Biol. **12:** 2749–2757.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

JEFFREYS, A. J., and S. D. J. PENA, 1993 Brief introduction to human DNA fingerprinting, pp. 1–20 in *DNA Fingerprinting: State of the Science,* edited by S. D. J. PENA, R. CHAKRABORTY, J. T. EPPLEN and A. J. JEFFREYS. Birkhäuser Verlag: Basel.

JEFFREYS, A. J., N. J. ROYLE, V. WILSON and Z. WONG, 1988 Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. Nature **322:** 278–281.

KELLY, R., M. GIBBS, A. COLLICK and A. J. JEFFREYS, 1991 Spontaneous mutation at the hypervariable mouse minisatellites locus Ms6-hm: flanking DNA sequence and analysis of germline and early somatic events. Proc. R. Soc. Lond. Ser. B **245:** 235–245.

KUNKEL, T. A., 1993 Slippery DNA and diseases. Nature **365:** 207–208.

LEVINSON, G., and G. A. GUTMAN, 1987 Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. **4:** 203–221.

MEYER, A., T. D. KOCHER, P. BASASIBWAKI and A. C. WILSON, 1990 Monophyletic origin of Lake Victoria cichlid fishes suggested by mitochondrial DNA sequences. Nature **347:** 550–553.

MORAN, P. A. P., 1975 Wandering distributions and the electrophoretic profile. Theor. Popul. Biol. **8:** 318–330.

NEI, M., 1972 Genetic distance between populations. Am. Nat. **106:** 283–292.

OHTA, T., and K. KIMURA, 1973 The model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic population. Genet. Res. **22:** 201–204.

QUELLER, D. C., J. E. STRASSMANN and R. H. COLIN, 1993 Microsatellites and kinship. Tree **8:** 285–288.

SCHLÖTTERER, C., and D. TAUTZ, 1992 Slippage synthesis of simple sequence DNA. Nucleic Acids Res. **20:** 211–215.

SHRIVER, M. D., L. JIN, R. CHAKRABORTY and E. BOERWINKLE, 1993 VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics **134:** 983–993.

SLATKIN, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. Genetics **139:** 457–462.

STEPHENS, J. C., D. A. GILBERT, N. YUHKI and S. J. O'BRIEN, 1992 Estimation of heterozygosity for single-probe multilocus DNA fingerprints. Mol. Biol. Evol. **9:** 729–743.

STRAND, M., T. A. PROLLA, R. M. LISKAY and T. D. PETES, 1993 Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature **365:** 274–276.

TAJIMA, F., and N. TAKEZAKI, 1994 Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. Mol. Biol. Evol. **11:** 278–286.

TAUTZ, D., 1993 Notes on the defunction and nomenclature of tandemly repetitive DNA sequences, pp. 21–28 in *DNA Fingerprinting: State of the Science,* edited by S. D. J. PENA, R. CHAKRABORTY, J. T. EPPLEN and A. J. JEFFREYS. Birkhäuser Verlag: Basel.

TODD, J. A., T. J. AITMAN, R. J. CORNALL, S. GHOSH, J. R. S. HALL *et al.*, 1991 Genetic analysis of autoimmune type 1 diabetes mellitus in mice. Nature **351:** 542–547.

VALDES, A. M., M. SLATKIN and N. B. FREIMER, 1993 Allele frequencies at microsatellite loci: the stepwise mutatin model revisited. Genetics **133:** 737–749.

WALSH, J. B., 1987 Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. Genetics **115:** 553–567.

WEHRHAHN, C., 1975 The evolution of selectively similar electrophoretically detectable alleles in finite natural populations. Genetics **80:** 375–394.

WEIR, B. S., A. H. D. BROWN and D. R. MARSHALL, 1976 Testing for selective neutrality of electrophoretically detectable protein polymorphisms. Genetics **84:** 639–659.

Communicating editor: W. J. EWENS

APPENDIX

To derive the expectation of D_1 as a function of time, define E_g as the single generation expectation operator. Then write

$$E_g[D_1(t)] = (2N)^{-2} E_g[\sum (i - i')^2 n_i(t)\, n_{i'}(t)], \quad \text{(A1)}$$

where the prime indicates the second population, and the sum is over all i, i'. Following MORAN (1975), denote $(2N)^{-1} \Sigma_i\, in_i(t)$ as $M_1(t)$ and $(2N)^{-1} \Sigma_i\, i^2 n_i(t)$ as $M_2(t)$. Then, add and subtract squared means to get,

$$E_g[D_1(t)] = E_g[M_2(t) - M_1(t)^2] + E_g[M_2'(t) - M_1'(t)^2] - 2(2N)^{-2} E_g \times [\sum ii' n_i(t)\, n_i'(t)] + E_g M_1(t)^2 + E_g M_1'(t)^2 \quad \text{(A2)}$$

$$= E_g[V(t) + V'(t)] + E_g[(M_1(t) - M_1'(t))^2], \quad \text{(A3)}$$

where $V(t)$ is the variance in allelic repeat number at time t. Then substitute MORAN's (1975) equations for the sampling and mutation process (his Equations 6 and 7) to get

$$E_g[D_1(t)] = (1 - \tfrac{1}{2}N)(M_2(t-1) - M_1(t-1)^2) + (1 - \tfrac{1}{2}N)\mu + (1 - \tfrac{1}{2}N)(M_2'(t-1) - M_1'^2(t-1)) + (1 - \tfrac{1}{2}N)\mu + (1 - \tfrac{1}{2}N) M_1 \times (t-1)^2 + \frac{1}{2N} M_2(t-1) + \frac{\mu}{2N} + (1 - \tfrac{1}{2}N) M_1'(t-1)^2 + \frac{1}{2N} M_2'(t-1) + \frac{\mu}{2N} - 2M_1(t-1) M_1'(t-1). \quad \text{(A4)}$$

Rewriting in terms of the variances in the previous generation, we have

$$E_g[D_1(t)] = (1 - \tfrac{1}{2}N)(V(t-1) + V'(t-1)) + 2(1 - \tfrac{1}{2}N)\mu + (1 - \tfrac{1}{2}N) M_1(t-1)^2 + \frac{1}{2N} M_2(t-1) + \frac{\mu}{2N} + (1 - \tfrac{1}{2}N) \times M_1'(t-1)^2 + \frac{1}{2N} M_2'(t-1) + \frac{\mu}{2N} - 2M_1(t-1) M_1'(t-1), \quad \text{(A5)}$$

which simplifies to

$$E_g[D_1(t)] = V'(t-1) + V(t-1) + 2\mu + M_1(t-1)^2 + M_1'(t-1)^2 - 2M_1(t-1) M_1'(t-1). \quad \text{(A6)}$$

Then, noting that $D_1(t-1) = V(t-1) + V'(t-1) + (M_1(t-1) - M_1'(t-1))^2$, we have

$$E_g[D_1(t)] = D_1(t-1) + 2\mu. \quad \text{(A7)}$$

Now, define E as expectation over multiple generations. Then, the average squared difference between alleles drawn from populations isolated τ generations ago is given by

$$E[D_1(t)] = D_0(0) + \tau 2\mu. \quad \text{(A8)}$$

Assuming the population was at equilibrium when separation occurred, we have,

$$E[D_1(t)] = 2(2N - 1)\mu + \tau 2\mu. \quad \text{(A9)}$$